Multichannel Acoustic Echo Cancelation in Multiparty Spatial Audio Conferencing With Constrained Kalman Filtering

Zhengyou Zhang, Qin Cai, Jay Stokes

Communications and Collaboration Systems
Microsoft Research, Redmond, WA, USA

zhang@microsoft.com

Outline

- Motivations
- Audio spatialization
- Two approaches to multichannel AEC
- Constrained Kalman filtering
- Experimental results

Motivation

- Current audio conferencing systems
  - Monaural → adequate for 1-to-1
  - Poor when #people > 2
- Why poor?
  - All the voice streams are intermixed into a single one
  - Huge cognitive load: *Do 2 things simultaneously*
    - *Associate voice signals to the speaker*
    - *Comprehend what is being discussed*

Solutions

- Video conferencing
- Spatial audio conferencing
- Spatial audio + Video
- Immersive conferencing

Solutions

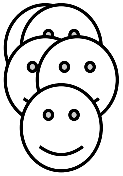
- Video conferencing
- Spatial audio conferencing
- Spatial audio + Video
- Immersive conferencing

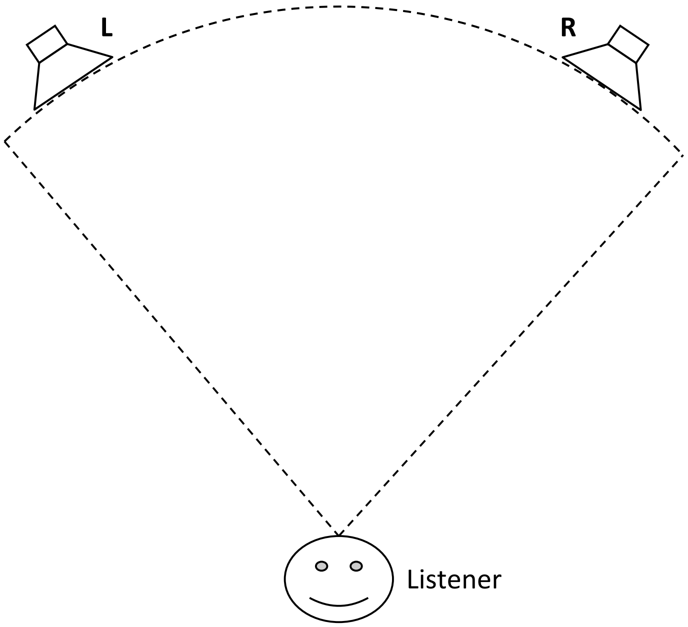
Benefits of Spatial Audio

- Human's cocktail party effect
  - Selective attention
  - Only spend effort on comprehension
- Brain rejects incoherent signals at two ears
  - Reverberation & noise are disregarded (not for mono!)
- Benefits:
  - Memory, Focal Assurance, Perceived Comprehension, Listener's Preference
  - http://msrweb/users/zhang/ThinkWeekPapers/Spatial%20audio%20conferencing.doc

Multiparty Spatial Audio Conferencing

- Virtual seating





Audio Spatialization

- Delay and Gain Modulation
  - Delay

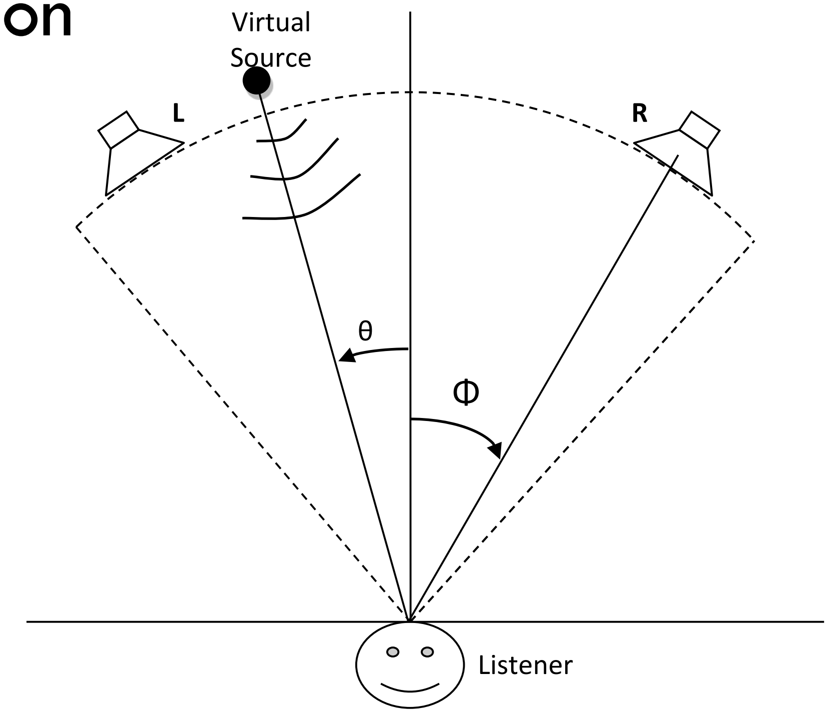
    $\Delta_R = D - D \cos(\lambda(\Phi - \Theta))$

    $\Delta_L = D - D \cos(\lambda(\Phi + \Theta))$

    $D=0.45ms \quad 1 \leq \lambda \leq \pi/(2\Phi)$
  - Gain

    $G_R = \cos(\lambda(\Phi - \Theta)/2)$

    $G_L = \cos(\lambda(\Phi + \Theta)/2)$

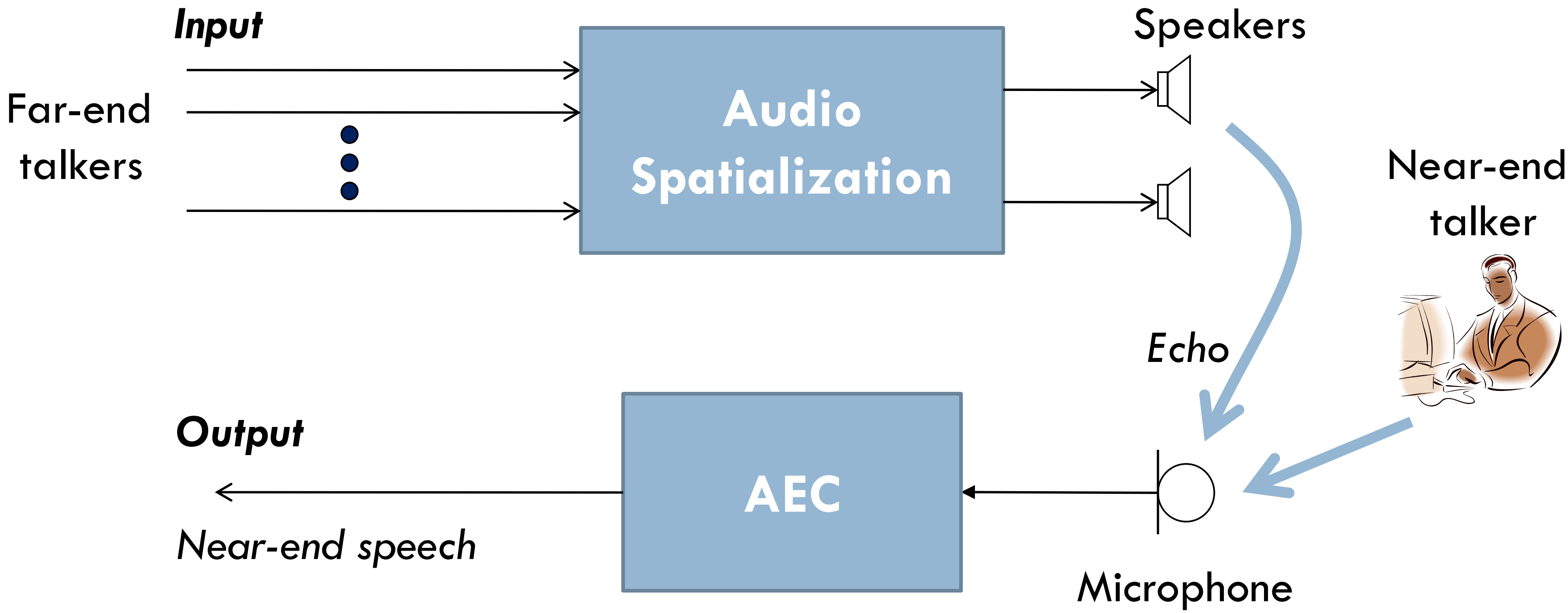


- Example:

  4 remote participants

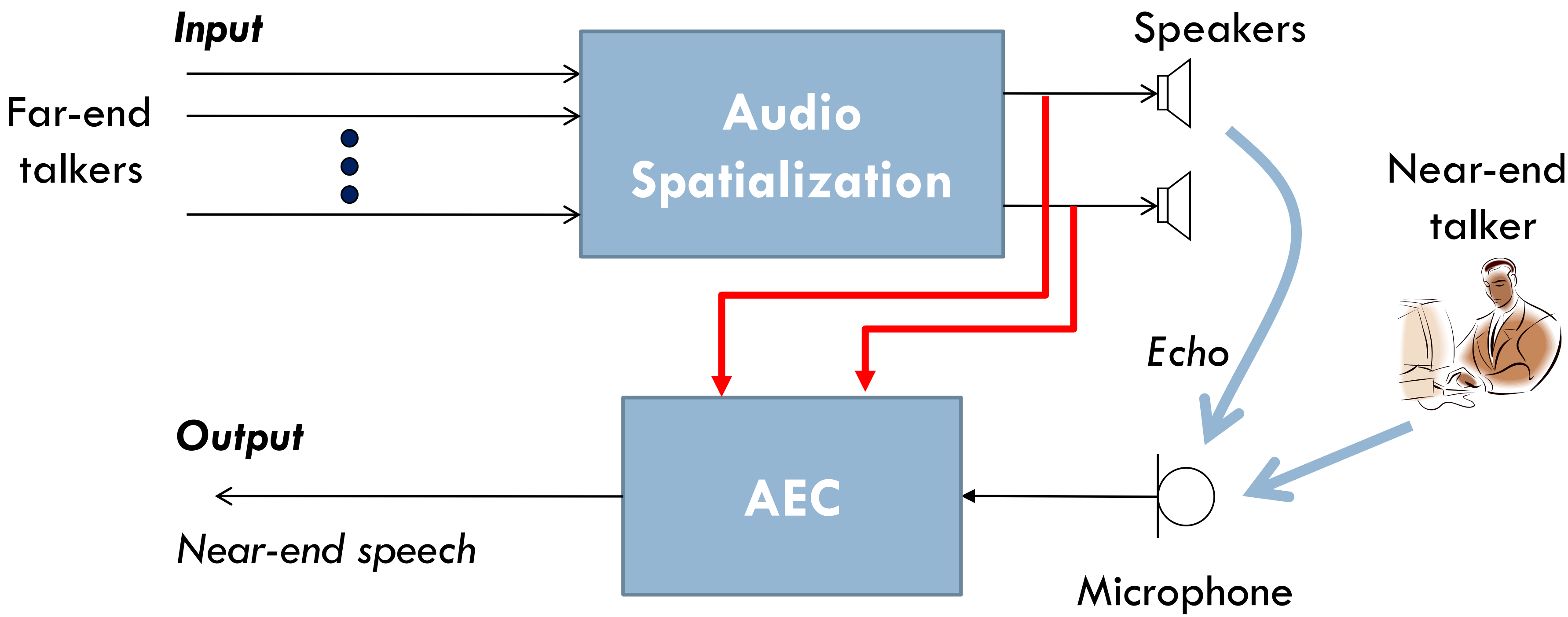
  [Traditional](short) ([short]) [Spatial] ([short])

Multichannel AEC



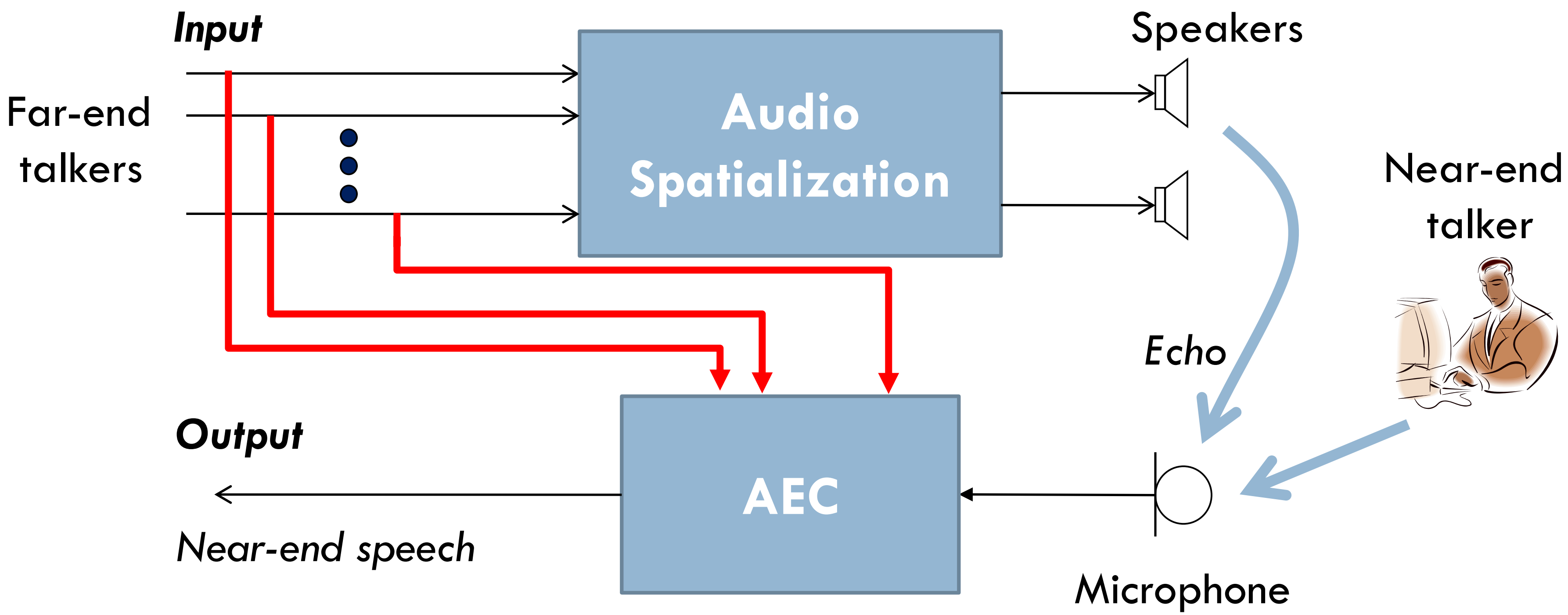
- **Question:** *Which reference signals to use?*

Approach 1: Use Speaker Signals



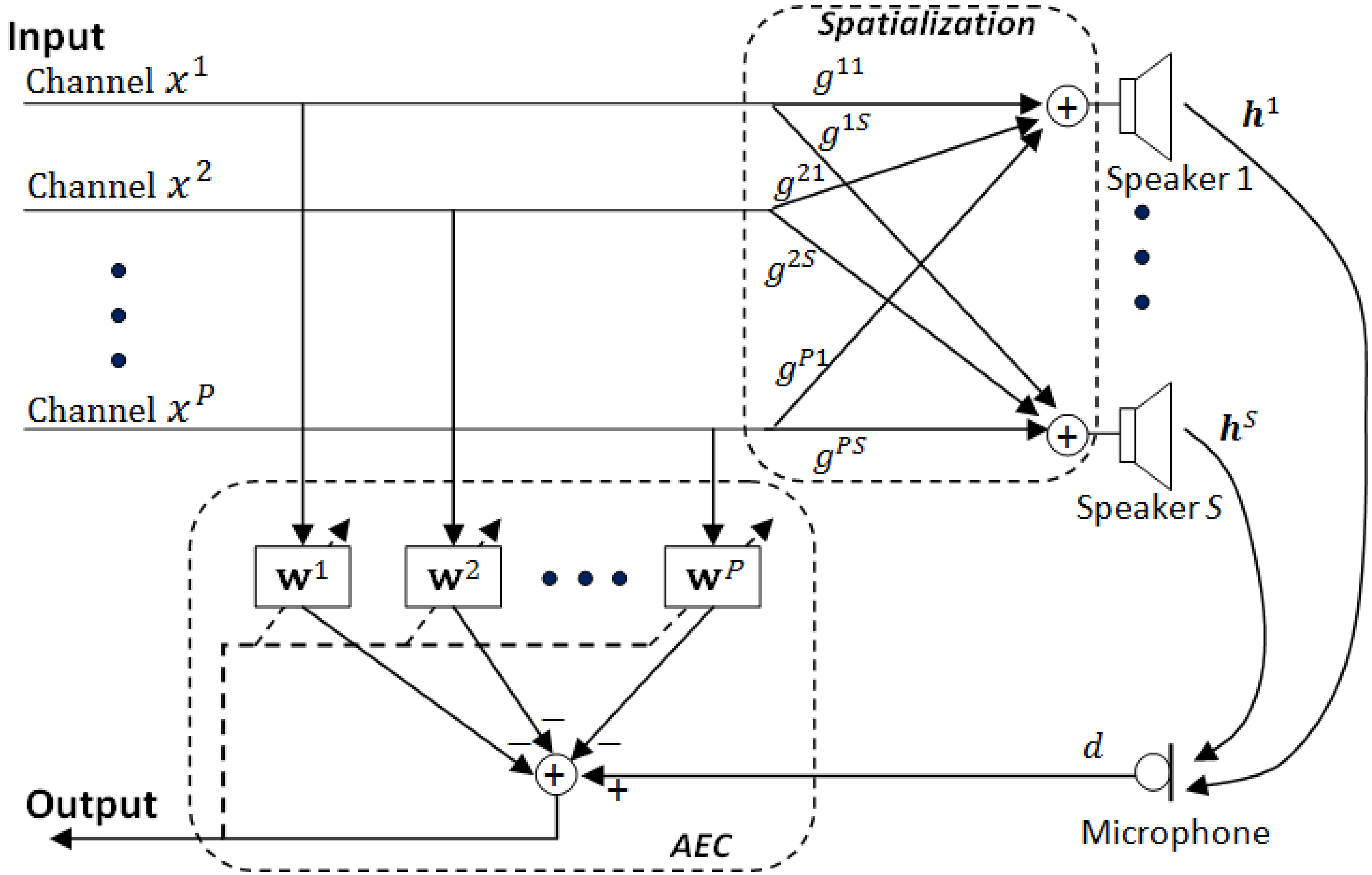
- Possible problem:
  - Correlation between speaker signals

Approach 2: Use Far-End Channels



- Cancel each individual far-end speech
- **Our solution:** Constrained Kalman filtering

Multichannel AEC: Diagram



Problem Statement

- Remote channels: $\{X^i \mid i = 1, \dots, P\}$
- Spatialization on S speakers: $Y^s = \sum_{i=1}^{P} G^{is} X^i$
- Speaker's room response: L-tap filter

$$\boldsymbol{H}_t^s = [H_t^s, H_{t-1}^s, \dots, H_{t-L-1}^s]^T$$

- Microphone input: Echo

$$D_t = \sum_{i=1}^{P} \sum_{s=1}^{S} G^{is} \left(H_t^s X_t^i + H_{t-1}^s X_{t-1}^i + \cdots + H_{t-L-1}^s X_{t-L-1}^i \right) = \sum_{i=1}^{P} \sum_{s=1}^{S} G^{is} \, \boldsymbol{H}_t^{s\,T} \boldsymbol{X}_t^i$$

Problem Statement (cont'd)

- Determine the echo cancellers:
  - one per remote channel i: L-tap filter

$$\mathbf{W}_t^i = \left[W_t^i, W_{t-1}^i, \ldots, W_{t-L-1}^i\right]^T$$

  such that echo is cancelled, i.e.,

$$D_t - \sum_{i=1}^{P} \mathbf{W}_t^{i^T} \mathbf{X}_t^i = 0$$

- Constraint: W_t^i's are not mutually independent

$$D_t = \sum_{i=1}^{P} \sum_{s=1}^{S} G^{is} \mathbf{H}_t^{s^T} \mathbf{X}_t^i \quad \Longrightarrow \quad \mathbf{W}_t^i = \sum_{s=1}^{S} G^{is} \mathbf{H}_t^s$$

Constrained Kalman Filtering

- State Vector: Echo cancellers + Speaker RIR filters

$$S_t = \left[W_t^{1^T}, \dots, W_t^{P^T}, H_t^{1^T}, \dots, H_t^{S^T}\right]^T$$

- System equation: $S_t = S_{t-1} + n_t$
- Observation equation: $D_t = A_t^T S_t + v_t$ with $A_t = \left[X_t^{1^T}, \dots, X_t^{P^T}, 0^{1^T}, \dots, 0^{S^T}\right]^T$
- Constraint: $CS_t = 0$ with

$$C = \begin{bmatrix} -1 & & & G^{11}1 & \cdots & G^{1S}1 \\ & \ddots & & \vdots & \ddots & \vdots \\ & & -1 & G^{P1}1 & \cdots & G^{PS}1 \end{bmatrix}$$

- New observation equation: observation + constraint

$Y_t = B_t S_t + v_t$ with $Y_t = [D_t, 0, \dots, 0]^T$ $B_t = \begin{bmatrix} A_t^T \\ C \end{bmatrix}$ $v_t = \begin{bmatrix} v_t \\ u_t \end{bmatrix}$

Constrained Kalman Filtering (cont'd)

- Assumptions

$$E[\boldsymbol{n}_t] = \mathbf{0} \qquad E[\boldsymbol{n}_t \boldsymbol{n}_t^T] = \boldsymbol{Q}_t$$

$$E[\boldsymbol{v}_t] = \mathbf{0} \qquad E[\boldsymbol{v}_t \boldsymbol{v}_t^T] = \boldsymbol{R}_t = \begin{bmatrix} \sigma_t^2 & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{\Lambda}_t \end{bmatrix}$$

Tuning parameter to control how hard the constraint be satisfied

- Equations

$$\boldsymbol{S}_t^- = \boldsymbol{S}_{t-1}$$

$$\boldsymbol{P}_t^- = \boldsymbol{P}_{t-1} + \boldsymbol{Q}_t$$

$$\boldsymbol{K}_t = \boldsymbol{P}_t^- \boldsymbol{B}_t^H (\boldsymbol{B}_t \boldsymbol{P}_t^- \boldsymbol{B}_t^H + \boldsymbol{R}_t)^{-1}$$

$$\boldsymbol{S}_t = \boldsymbol{S}_t^- + \boldsymbol{K}_t (\boldsymbol{Y}_t - \boldsymbol{B}_t \boldsymbol{S}_t^-)$$

$$\boldsymbol{P}_t = (\mathrm{I} - \boldsymbol{K}_t \boldsymbol{B}_t) \boldsymbol{P}_t^-$$

Benefits of Constrained KF

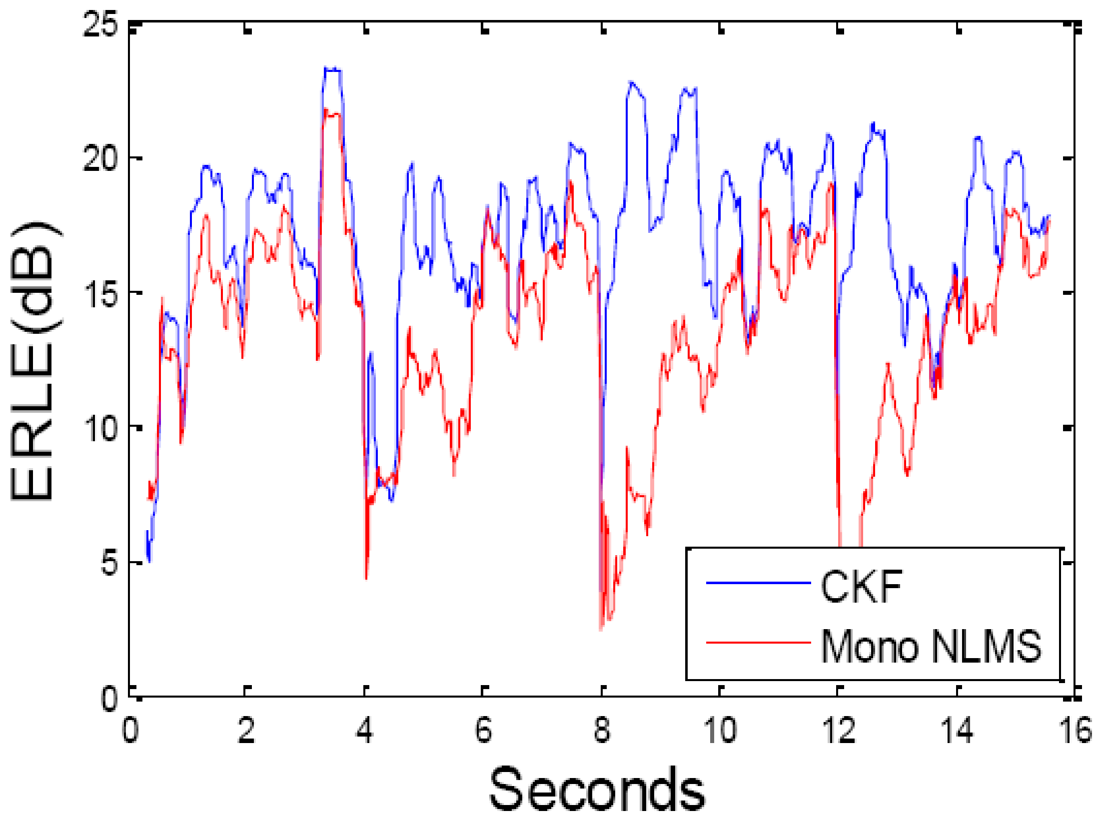
- The constraint is taken care of automatically, and can be imposed *with varying degrees*.
- All channels are taken into account simultaneously. →Overlapping far-end talking is not an issue
- The AEC for each channel is updated continuously because of the constraint, even if it is inactive. →AEC's are always up to date
- Ambient noise can be time varying. → Use a separate noise tracker

Comparison with Prior Art

- T.N. Yensen, R.A. Goubran, and I. Lambadaris, "Synthetic Stereo Acoustic Echo Cancellation Structure for Multiple Participant VoIP Conferences", IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 2, pp. 168-174, Feb. 2001.
- Same: One canceller per remote channel
- Differences:
  - Constrained vs. independent cancellers
    - Additional canceller is initiated before being active
    - A canceller is updated even if it is not active
  - Frequency vs. time domain
  - KF (RLS) vs. NLMS

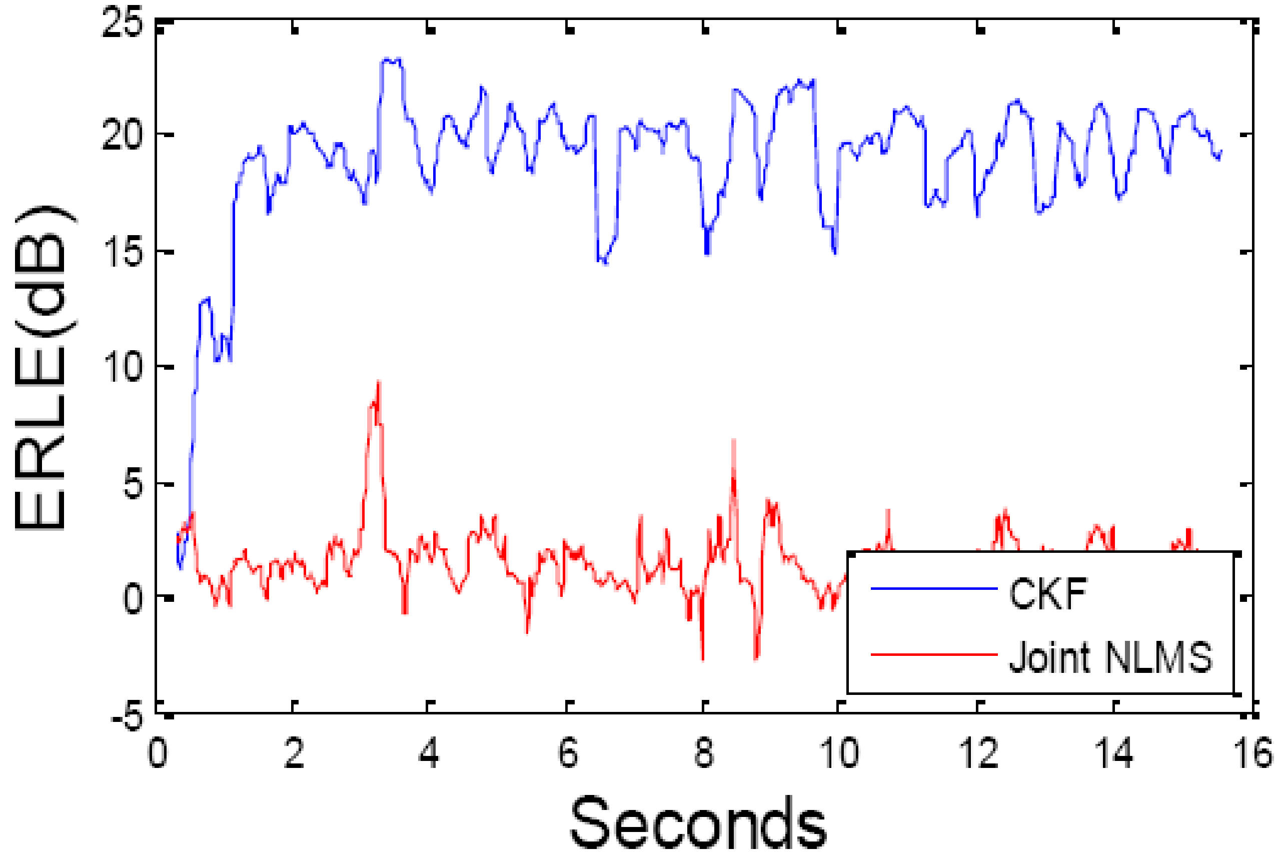
Experimental Results

- Simulation setup:
  - 4 remote talkers at [-30°, 30°, 0°, -45°]
  - Each talks for 4s
  - Noise: -20dB
  - Fixed RIR
- Comparison
  - Constrained KF
  - Multiple mono NLMS



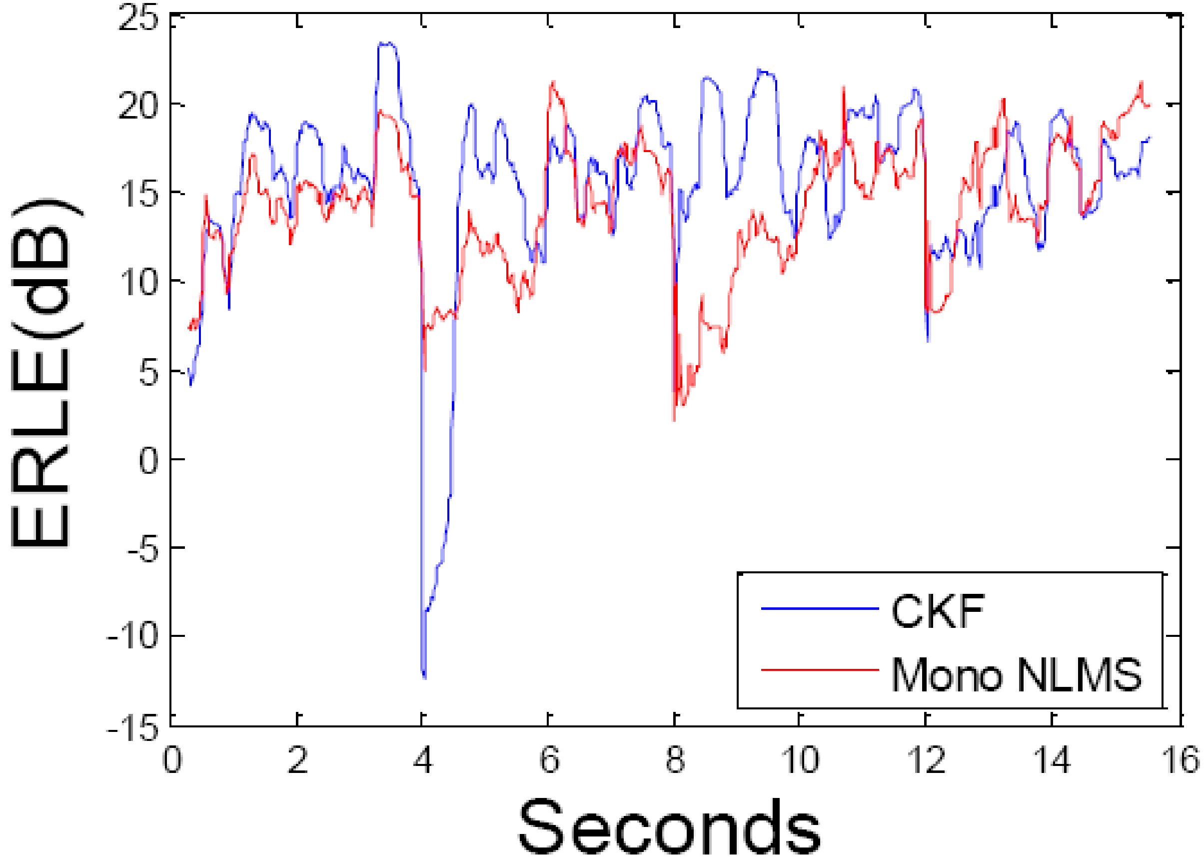
Experiment: Overlapping Talkers

- Two simultaneous remote talkers



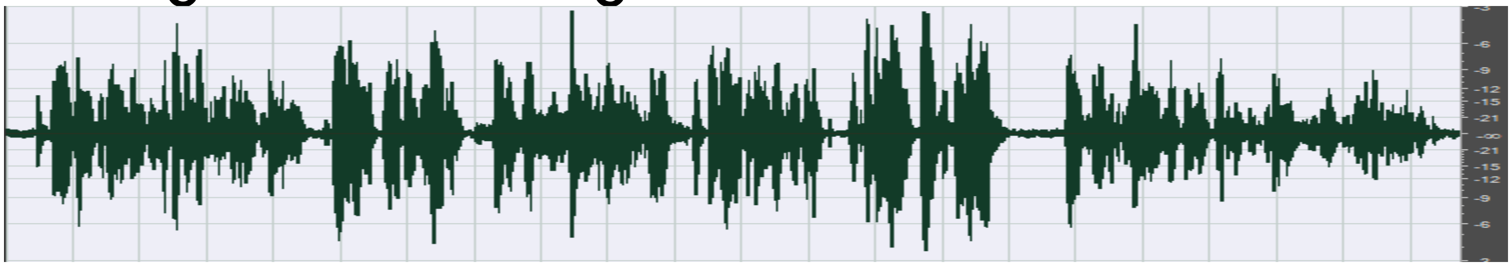
Experiment: Changing RIR

- -30dB change in RIR every 0.5 seconds

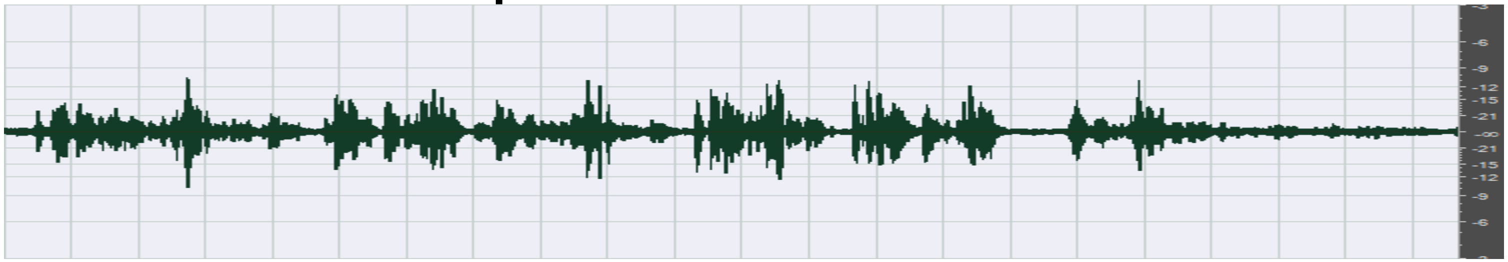


Experiment with real data

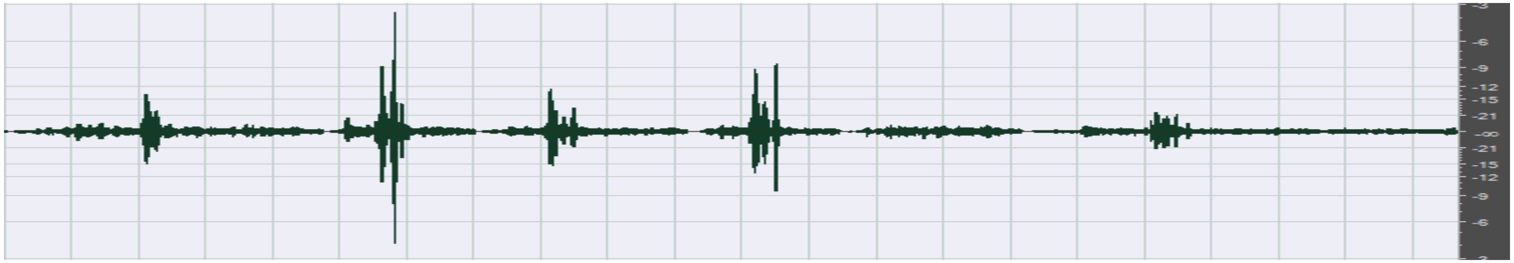
- Original recording with near-end talker



- AEC with multiple mono NLMSs



- AEC with CKF



Conclusions

- Constrained Kalman Filtering for multichannel AEC
- Outperform over multiple independent mono AECs
  - Additional canceller is initiated before being active
  - A canceller is updated even if it is not active
- Naturally works with multiple simultaneous remote talkers without resort to channel switching

Thank you !

Q & A